

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-239799

(43)Date of publication of application : 12.09.1995

(51)Int.Cl. G06F 11/20  
G06F 3/06  
G06F 13/00

(21)Application number : 06-301109

(71)Applicant : INTERNATL BUSINESS MACH  
CORP <IBM>

(22)Date of filing : 05.12.1994

(72)Inventor : KERN ROBERT F  
MICKA WILLIAM F  
MIKKELSEN CLAUS W  
PAULSEN MICHAEL A  
SHOMLER ROBERT WESLEY

(30)Priority

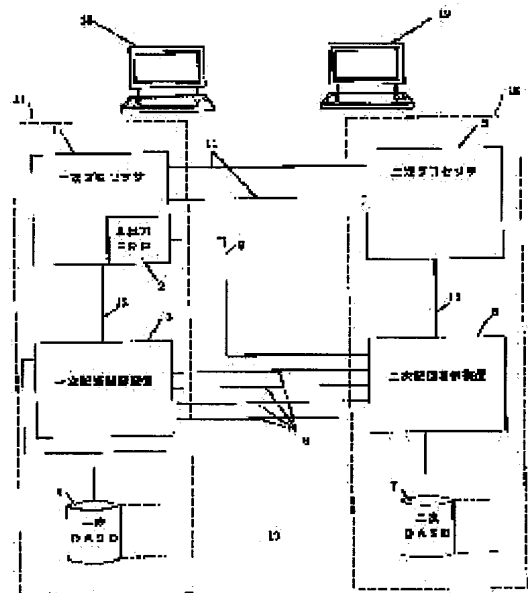
Priority number : 94 199448 Priority date : 22.02.1994 Priority country : US

## (54) METHOD FOR PROVIDING REMOTE DATA SHADOWING AND REMOTE DATA DUPLEX SYSTEM

(57)Abstract:

**PURPOSE:** To provide a remote data shadowing system which synchronously performs a real-time disaster recovery in a memory location base with a secondary side placed at a separated location from a primary one.

**CONSTITUTION:** An error recovery program on a primary side 14 executes error recovery procedures and then stops the application in its execution mode to notify primary and secondary sides 14 and 15 that dual pair of faults take place. The error recovery program discriminates the cause of the dual pair of faults and restarts a dual mode if an error recovery is normally performed. When the error recovery program can not normally execute the error recovery, the update of writing on the primary side is further inhibited and an error message is sent to the operators of both primary and secondary sides 14 and 15.



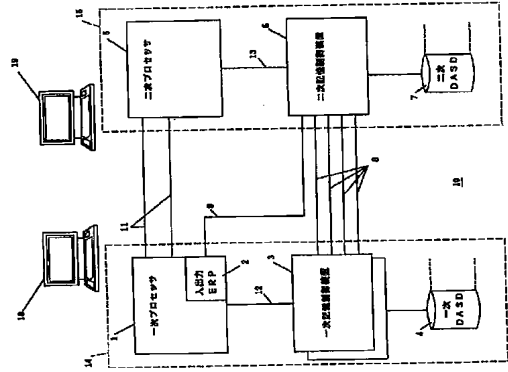
識別記号	社内整理番号	FI	技術表示箇所
(51) Int.Cl. G 0 6 F	11/20		
	3/06		
	301 P		
	13/00		
(31) 優先権主張番号	特願平6-301109	(71) 出願人	380009531
(32) 優先日	平成 6 年 (1994) 12 月 5 日		インターナショナル・ビジネス・マシーンズ・コーポレーション
(33) 優先権主張国	米国 (US)		INTERNATIONAL BUSIN ESS MACHINES CORPO RATION アメリカ合衆国 0604、ニューヨーク州 アーモニー (所在地なし)
		(72) 発明者	ロバート・フレデリック・カーン アメリカ合衆国 85730 アリソナ州ツー ン イースト・カレヒコ・ストリート 8338
		(74) 代理人	弁理士 台田 義 (外 2 名) 最終頁に続く

(54) 【発明の名称】  
 遠隔データ・シャドーイングを提供する方法および遠隔データ二重化システム

(57) 【要約】

【目的】 二次側が一次側から離れた位置に配置され、同期的に記憶域ベースのリアルタイム災害復旧を行う遠隔データ・シャドレーイング・システムを提供する。

【解説】 一次側のエラー回復プログラムは、エラー一回復手順を実行し、二重側の障害が発生したときを一次側と二次側の両方向に通知すると、そこで実行中のアプリケーションに停止する。このエラー回復プログラムはリケーションの理由を判断し、エラー回復が正常に行われ二重側の障害の原因を判断し、エラー回復が正常に行われた場合は、二重モードが再開される。エラー回復プログラムがエラー一回復を正常に実行できない場合は、一次側で二重モード更新がさらに発生し、一次側と二次側両方のオペレータに対してエラー・メッセージが送達される。



(2)

1

【匪類の食器茶器】

【請求項1】一例側が一次記憶サブシステムに接続された二次プロセッサを含み、前記二次プロセッサが一次記憶サブシステムから入力されたデータを更新する少なくとも一つのアプローチをその中で実行し、前記二次プロセッサがそこで実行される入出力エンバードプログラム（入出力ERP）をさらに有し、二次側が二次記憶サブシステムに接続された二次プロセッサを含み、前記二次側が入出力変換更新を同期的にシャドウイングし、前記二次記憶サブシステムが入出力ERPと通信する。前記二次記憶サブシステムは入出力ERPと通信する。前記二次・シャドウイングを提供する方法において、

(a) 前記一次記憶サブシステムから前記二次記憶サブシステムに入出力書込み更新を送るステップと、  
(b) 前記一次および二次記憶サブシステムに二重対称性が確立されるかどうかを判断するステップと、  
(c) 一次記憶サブシステムから一次プロセッサに障害が発生した二重対称性を報告するステップと、  
(d) 二重対称の障害が発生した場合に、少なくとも1つのアプリケーションから出力ERPに一次プロセッサの制御権を移転するステップと、  
(e) 二重対称の障害の原因を判断するステップと、  
(f) 入出力ERPから二次記憶サブシステムに二重対称の障害発生メッセージを送るステップとを含む方法。  
【請求項2】二重対称性が設定されるが回復されなかった場合に前記一次記憶サブシステムが更新の書き込みを続行することを特徴とする、請求項1に記載の遠隔データベースシステムを提供する方法。

【請求項3】 原告が発生した二重対を報告する際に前記一次記憶サブシステムが前記一次プロセスに出入力状況としてチャネル終了/装置終了/装置チェックを報告することがステップ(c)に含まれることを特徴とする、請求項2に記載の選路データ・シャドイングを提供する方法。

【請求項4】 入力ERPが一次記憶サブシステムにセ  
ンス入出力を出すことがステップ(d)に含まれること  
を特徴とする、請求項3に記載の遠隔データ・シャドー  
イングを提供する方法。

【請求項5】二重対の障害が発生した場合に入出力ER  
Pによるエラー回復を実行し、二次側でデータ保全本性を  
維持するためのステップ(g)をさらに含むことを特徴

とすると、請求項4に記載の遠隔データ・シャドーイングを提供する方法。

【請求項6】エラー回復が正常に完了した後で入出力EPPから少なくとも1つのアプリケーションに制御権を

返すためのステップ (h) をさらに含むことを特徴とする、請求項5に記載の遠隔データ・シャドーイングを提供する方法。

【請求項7】エラー回復が正常に完了できなかった場合に、入出力ERPを介して二次側に障害発生同期遠隔コピ

[illegible]

特開平7-239799

2

一・メッセージを送るためのステップ (i) をさらに含むことを特徴とする、請求項6に記載の遠隔データ・シヤドーイングを提供する方法。

【請求項8】ステップ（i）によって、一次記憶サブシステムへの後続の入出力番込みがさらに防止されることを特徴とする、請求項7に記載の遠隔データ・シャドーイングを提供する方法。

【請求項9】一次プロセッサと二次プロセッサとの間で、エラー回復のための通信が行われることを特徴とする、請求項5に記載の遠隔データ・シャドーイングを提供する方法。

【請求項10】一次記憶サブシステムと二次記憶サブシステムとの間でエラー回復のための通信が行われることを特徴とする、請求項5に記載の遠隔データ・シャドーイングを提供する方法。

【語彙項目11】一次側と二次側を有し、データ・バックアップのために二次側が一次側から書き込み更新を受け取り、一次側を使用不能にする故障が発生した場合に、二次側がリアルタイムのデータ可用性を提供できるように、二次側が一次側から十分離れた位置に配置される遠隔データ二重化システムにおいて、遠隔データ二重化システムが、

雷込み人出力更新を生成する少なくとも1つのアプリケーションを実行する、一次側の一次プロセスと、雷込み人出力更新を受け取って格納するための一次記憶サブシステムと、

データ・リンクによって一次プロセスに連結される、一次側の一次プロセスと、

一次プロセッサに接続され、通信リンクを介してさらにバック記憶サブシステムに接続され、書き込み出力更新の受け取ったとき特定の記憶アドレスに書き込み出力更新が送られる。二次間の二次記憶サブシステムと、一次プロセッサ内にあって、二次記憶サブシステムに連結され、二重対の故障が発生した場合にエラー回復を行うためのエラー回復手段とを含む遠隔データ二重化システム。

【請求項12】エラー回復手段が、一次プロセスと二次側の両方に二重対障害発生モードを報告することを特徴とする、請求項11に記載の遠隔データ二重化システム。

【請求項13】エラー回復手段が、二重対の障害の原因を判別しながら、一次プロセスで実行中のアプリケーションを静止することを特徴とする、請求項11に記載の遠隔データ二重化システム。

【請求項14】エラー回復手段が、一次記憶サブシステムにセンス入出力を出すことを特徴とする、請求項13に記載の遠隔データ二重化システム。

【請求項15】エラー回復手段がエラー回復を完了し、二重モードが再確立された後、エラー回復手段が一次ブ



を判断することを含む。二重対の降格は、一次記憶サブシステムから一次プロセッサに報告される。一次プロセッサは、二重対の降格が発生したときに少なくとも1つのアプリケーションから入出力ERPに制御権を移転し、入出力ERPは、二重対の降格の原因を判別し、二重対降格発生メッセージを二次記憶サブシステムに送信する。

【0012】本発明の他の実施例では、遠隔データ二重化システムは一次側と二次側を含み、二次側は、データバックアップのために一次側に呼ばれる。二次側は、データを更新する必要があるときに一次側に呼び込まれ、更新を受け付ける。二次側は、一度使用不能になる必要がある場合、一次側から十分離れた位置に配置される。一次側の一次プロセッサは、書き込み出力更新の原因となる少なくともひとつのアプリケーションを実行し、一次記憶装置を介して一次記憶サブシステムに接続され、さらに通信リンクを通じて、データ・リンクによって一次プロセッサに接続されている。二次プロセッサは、二次記憶サブシステムに接続され、データ・リンクによって一次プロセッサに接続される。同じく一次側に位置する一次記憶サブシステムを介して一次記憶サブシステムに接続され、さらに通信リンクを通じて、二次プロセッサに接続され、さらに通信リンクを介して一次記憶サブシステムに接続され、二次記憶サブシステムは、受け取った順に書き込み出力更新が二次記憶サブシステムに書き込まれるような、書き込み出力更新のバックアップの記憶域ベースの二重化を提供する。一次プロセッサの出力およびエラー回復は、二次記憶サブシステムに提供される。

【0013】本発明の上記およびその他の目的、特徴、および利点は、添付図面に図示する本発明の実施例に関する以下の詳細な説明から明らかになるだろう。

【0014】一般的なデータ処理システムは、データを多重実行して操作し、データ機能管理サブシステム/多重仮想記憶システム (DFSMS/MVS) ソフトウェアなどを実行するために、IBMシステム/360またはIBMシステム/370プロセッサなどのホスト・プロセッサの形態で、少なくとも1台のIBM 3990記憶制御装置がそれに接続され、その記憶制御装置が、メモリ制御装置と、それに組み込まれた1つまたは複数のタイプのキャッシュ・メモリとを場合合わせ、さらに記憶制御装置は、IBM 3380または3390 DASDなどの1群の直接アクセス記憶装置 (DASD) に接続されている。ホスト・プロセッサが、実質的な計算能力を提供するのに対し、記憶制御装置は、大規模データベースを効率よく伝送し、ステージ (stage) / デス・ステージ (destage) し、変換し、全般的にアクセスするのに必要と階層化を提供する。

【0015】一般的なデータ処理システムの災害復旧保護では、一次DASDに格納した一次データを二次側または遠隔地でバックアップする必要がある。一次側と二

この場合、ユーザが受け入れられる危険のレベルによって、数キロメートルから数千キロメートルの範囲で可変である。二次側または遠隔側は、バックアップ・データ・コピーを保持するだけでなく、一次システムが使用不能になった場合に一次システムの処理を引継ぎ、その理由を一次システムに通知するだけならば、二次側と二次側には一次および二次DASDストリング間の両方にデータを書き込まないためである。むしろ、一次側と二次側に接続された一次DASDストリングに一次側制御装置に接続された二次DASDストリングには一次データが格納されるのに対し、二次側制御装置に接続された二次DASDストリングには二次データが格納されるのである。

【0016】二次側は、一次側から十分離れている必要があるため、一次データをリアルタイムでバックアップできず、二次側は、一次データが更新されたときに、最小の遅延で一次データをバックアップする必要がある。しかも、二次側は、一次側で実行されるデータまたは更新生成するアプリケーション・プログラム（たとえば、IMSやDB2）を考慮せず、二次側に、一次データをバックアップしなければならない。二次側に要求される難しい課題は、二次データの順序が整合していないなければならないことである。つまり、二次データは一次データと同じ順序でコピーされない限り、二次データと整合性（順序整合性）の検証が難しい。二次データを要する順序である。例えば、整合性は、それらが1つのデータベースシステム内の複数のDASDを制御する記憶制御装置間の複製存在のためにさらに複雑になっている。順序整合性がないと、一次データと一致しない一次データが生成され、その結果、災害復旧が複雑する恐れがある。

【0017】遠隔データの二重化は、同期・非同期コピーで一つの体系的なカテゴリに分けられる。同期遠隔コピーでは、一次データを二次側に送り、一次DASDの出力ポートから二次DASDの入出力ポートにチャネル終了(CE)操作を終了する(一次ポートにチャネル終了(CE)装置終了(DE)を出力する)前このようなデータの出入りやり取りを必要とする。このため、同期コピーでは、二次側の確認を待っている間に一次DASDの出入力ポートが埋まる。一方、非同期コピーでは、一次側の出力が空になる。一次側の出力ポートとの距離に比例して長くなる。二次システムと二次システムとの距離に比例して長くなる(これは距離距離を数千キロメートル規模に制限する要素である)。しかし、同期コピーでは、システム・オーバーヘッドと比較的小さくして、順序が整合したデータを二次側に提供する。

【0018】非同速連環エビでは、二次側でデータがタ  
 ーに渡される前に一次DASDの入出力操作が完了する  
 (一次出力にチャネル終了(CE)と装置終了(DE)  
 を出力する)ため、一次側のアプリケーション・シ  
 ステムのパフォーマンスが向上する。このため、一次D  
 ASDの入出力応答時間は一次側までの距離に依存す

ず、一次側から数千キロメートル離れた遠隔地に二次側を設けることもできる。しかし、二次側で受け取ったデータは、一次側で受け取ったデータと一致しない場合がある。この原因は、データが一次側の順序整合性を確保するのに必要なシステム・オーバーヘッドが増加する。したがって、一次側で障害が発生すると、一次側と二次側との間で転送中のデータが一部消失する恐れがある。

【0019】同期データー・シャード・インデックス復旧のための同期リブライティング遠隔コピーでは、コピーされるデータのDASDボリューム1つのセグメントを形成する必要がある。さらに、このようなセグメントを形成させるには、各セグメントを構成するこれらのボリュームとそれらに対応する一次側の同等物（VOLSER）を識別する必要がある。必要なのは、二次側が一次側と「二重対（duplex pair）」を形成するの、1つまたは複数のデータセグメントと同期していない、つまり、「二重対の錯覚」が発生したときに二次側が再試行される間に、DASDの入出力が連続するため、非同期遠隔コピーにより同期遠隔コピーの方が、孩童誤差を認識しやすい。

[illegible]

【0020】しかし、二次DASDが存在しアクセス可能である状態では、二次側と一次側の接続を維持し、内容の同期と確保は保たれない。いくつかの理由から、二次側は一次側との同期性を失う場合もある。二次側が形成されたときと二次側は当初同期しておらず、初期データセットは一次側と二次側は当初同期に達する。一次側が二次側に更新されたデータを書き込めない場合、一次側は更新されたデータを保持することもある。この場合、更新されたデータを保持することが行われるように、二次側が更新された状態では二次DASDに更新された状態を書き込む。そのため、二次側が更新されるまで、一次側は露出状態では、

つまり、現行の災害保護コピーを「使わずに実行」を続ける。二重ガードが完了されても、二重側は直ちに同期状態になるわけではない。この時点で保護状態の更新を適用した後は、二重側は同期状態に戻る。一方側は、該データリユーザに関する中止コマンドを二重DASDに出すことにより、二重側の同期を喪失させることもできる。この中止コマンドが終了し、二重が再び確立され、保護状態の更新がコピーされたと、二重側は一次側と再度同期する。また、オンライン保守中も、同期を喪失させることがで

【0021】二次ポリユーラムが一次ポリユーラムと同様に、二次ポリユーラムは、二次システムの復旧時に一度と一回側アプリケーションの再開に使用することができる。一方、「一次側の非同期ポリユーラムは非同期ポリユーラムではない」として識別されなければならない。一次側の非同期ポリユーラムは、アプリケーション・アクセスを否定する（そのポリユーラムを強制的にオフラインにするか、そのVOLSE Rを変更する）ために非同期ポリユーラムを識別する必要があり、一次側ポリユーラムがアクセス不能になった場合がある。このため、二次側では、すべてのポリユーラムの同期サブシステム、すなわち、二次記憶制御装置CDASおよびサブシステム、また、一次側で抽出された例外によって一次出力経路を除く原因となるすべての条件を判断できるわけではない。たとえば、二次側が把握していない一次出力経路の状態に関するすべての前提情報が必要である。二次記憶サブシステム、二次記憶制御装置CDASと二次側ポリユーラムとの関係は、二次側がアクロスできない場合、一次側は二次側を排除することになる。この場合、この場合、二次側が同期状態を示し、一次側は、二次側が排除されたことを示す。

【0002】非同期二重対称ユーザが存在すること  
を、外通知音によって二側側に通知することができ、  
これは、ユーザ・システム管理機能を使用することで  
認識できる。一側側の入出力操作はチャネル終了、接続終  
了、接続チェック (CE/DE/UC) 状態を終了し、  
セクセンス・データがエラーの特徴を示す。このような形式  
の入出力形成の場合、エラー回復プログラム (ERP)  
がエラーを処理し、入出力の完了を二側側アプリケーション  
に通知する前に二側側メッセージを送る。この場合、ERPの二重対称メッセージを設  
置し、その情報を二側側に提供するのは、ユーザの責任  
である。一側側の代わり動作が可能になるよう二側側が  
順りにはユーザにオナラン接続し、アプリケーション  
Dがユーザにオナラン接続し、アプリケーション  
Dの制御のために非同期二重対称ユーザがオプション接続  
されていないことを確認するために、二重DASDサブ  
システムに接続された同相状況が検査される。この同相  
状況をすべてのERP二重対称ユーザ全体を示すクチャ  
ーが得られる。

【0023】ここで図1を参照して説明すると、図面1は一次元データは、一次元14と二次元1.5を有し、二次元1.5が一次元1から20キロメートル離れていて、炭素鋼システム14から14キロメートル離れたところにある。一次元14は、そこで実行されたシステム14に示されているアプリケーションと、システム14を入力および出力しているアプリケーションと、システム14（以下、入力ERP2という）の逆方向プログラム2（以下、出力ERP2という）のプロセスまたは一次プロセス1とを有するホスト・プロセスまたはDF SMS/MVSオペレーティング・ソフトウェアを含む。一次プロセス1は、IBMエンタープライズ・システム/9000（ES/9000）など

との間、または最悪の場合は、一次プロセッサ1と二次プロセッサ5との間のデータセキュリティを維持するために、新たな対等通信同期エラー回復を実行することができ

る。

【0026】図2および図3を参照して説明すると、同図にはエラー回復手順が示されている。図2のステップ201は、一次プロセッサ1で実行されるアプリケーション・プログラムが一次記憶制御装置3にデータ更新を送信することを含む。ステップ203では、そのデータ更新が一次DASD4に書き込まれ、そのデータ更新が二次記憶制御装置6にシャドーイングされる。ステップ205では、二重対の状態がチェックされ、一次側と二次側が同期しているかどうかが判別される。二重対の状態が同期状態になっている場合、ステップ207でデータ更新が二次DASD7に書き込まれ、一次プロセッサ1での処理は、そこで実行されるアプリケーション・プログラムを介して続行される。

【0027】二重対が「障害発生」状態になっている場合、ステップ209で一次記憶制御装置3は、二重対で中断または障害が発生していることを一次プロセッサ1に通知する。二重対は、通信リンク8による一次記憶制御装置3と二次記憶制御装置6との通信障害によって「障害発生」状態になる場合がある。あるいは、二重対は、一次サブシステムまたは二次サブシステムいずれかのエラーによって「障害発生」状態になる場合もある。

障害が通信リンク8で発生している場合、一次記憶制御装置3は、二次記憶制御装置6に直接、障害を連絡することができない。そこで、一次記憶制御装置3は、ステップ211で出力状況としてCE/DE/UCを一次プロセッサ1に返す。出力ERP2は、アプリケーション・プログラムを静止させ、書き込み出力操作を要求するアプリケーションに制御権を返す前に、エラー回復とデータセキュリティのためにステップ213で一次プロセッサ1の制御権を引き継ぐ。

【0028】図3は、出力ERP2が実行する諸ステップを表している。ステップ221で出力ERP2は一次記憶制御装置3にセンス入出力を出す。センス入出力操作は、入出力エタラの原因を記述する情報を返す。すなわち、このデータ記述情報は、具体的なエラーに関する記憶制御装置または二重対の操作に固有のものに化す。一次記憶制御装置3と二次記憶制御装置6との間の如等通信リンク8で障害が発生したことがデータ記述情報によって示された場合、ステップ223で出力ERP2は、一次記憶制御装置3および二次記憶制御装置6に対して、関係ボリュームを障害発生同期遠隔コピー状態に入れるように指示する記憶制御装置レベル入出力操作を出す。この二次記憶制御装置6は、複数のESCOリンク9または二重対間通信リンク11を介して出力ERP2から関係ボリュームの状態を受け取ることができ、その結果、二重対操作の現在の状況は、一次

ロセッサ1で実行されるアプリケーションとともに、一次プロセッサ1および二次プロセッサ5の両方で維持される。コンソール18および19は、それぞれ一次プロセッサ1および二次プロセッサ5からの情報をやりとりするために設けられ、入出力ERP4は、両方のコンソール18および19に状況情報を通知する。

【0029】一次記憶制御装置3および二次記憶制御装置6への障害発生同期遠隔コピー入出力操作が正常に完了すると、ステップ225でデータセキュリティが維持されてくる。このため、二次側15で復旧を試みると、二次記憶制御装置6は、「障害発生同期遠隔コピー」というマークを付けたボリュームを、データ回復手段（ボリューム1上のそのデータの状態を判別するための従来のデータベース・ログまたはジャーナル）によってそのボリュームのデータとその同期グループ内の他のデータとの同期が取られるまで使用できないものとして識別する。

【0030】ステップ227では、障害発生同期遠隔コピーの状況更新について一次記憶制御装置3と二次記憶制御装置6で行われた入出力操作の正常終了を入出力ERP2が受け取ったかどうかを判別するテストが行われる。正常終了すると、入出力ERP2は、ステップ229で一次プロセッサ1に制御権を返す。正常終了していない場合は、ステップ231で次のレベルの復旧通知が行われる。この通知には、障害発生ボリュームと、一次記憶制御装置3または二次記憶制御装置6のいずれかのそのボリュームの状況が正しくない可能性があることを、コンソール18を介してオペレータに通知することを含まれる。この通知は、そこで具体的なボリューム状況を示すために、コンソール19または共用DASDデータ・セットを介して次側15にシャドーイングされる。

【0031】ステップ233で、エラー・ログ記録データ・セットが更新される。この更新は、一次DASD4または他の記憶場所のいずれかに書き込まれ、二次側15にシャドーイングされる。このエラー回復処理が完了すると、入出力ERP2はステップ235で、障害発生書き込み入出力操作に関する「永続エタラ」を通知するアプリケーションに実行させるために、一次側アプリケーションの書き込み入出力操作に「永続エラー」を通知する。エラーが修正されると、ボリューム状態は、まず保留状態（変更データの再コピー）に回復し、次に全二重対に回復することができ、その後、二重対が再確立されると、データを二次DASD7に再適用することができ

る。

【0032】二重対を確立する場合、顧客の要求に応じて、ボリュームをCRITICALと識別することができ、CRITICALボリュームの場合、ある操作の結果、二重対の障害が発生すると、実際のエラー箇所とは無関係に、一次ボリュームの永続エラー障害が報告される。CRITICALの場合、障害発生対の一次DASD

406に書き込むとするその後のすべての試みは、永続エラーを受け取ることになり、対をなす二次ボリュームにシャドーイングできないデータは、その一次ボリュームに一切書き込まれなくなる。このため、必要であれば、一次側アプリケーションの処理および入出力データ操作との完全同期が可能になる。

【0033】その結果、本明細書に記載する災害復旧システム10では、入出力命令（チャネル・コマンド・ワード（CCW））を有する一次ホスト処理エタラ回復手順によって、二重対から障害発生二重対へ一次および二次同期遠隔コピー・ボリュームの状況を変更できるように外部同期遠隔コピーを取り入れ、それにより、複数タイプの一次および二次サブシステム・エタラの場合にデータ安全性を維持する。アプリケーション・ベースのパックアップではなく、データ更新がリアルタイムで書きされる記憶域ベースのバックアップが設けられている。また、災害復旧システム10は、（1）一次および二次記憶制御装置ボリューム状況更新、（2）オペレータ・メッセージまたはエラー・ログ記録共通データ・セットを介して具体的なボリューム更新状況に關して一次および二次ホスト・プロセスが通知すること、および（3）CRITICALボリューム識別などの、複数レベルの一次/二次状況更新を試み、ボリューム対が障害発生二重対になる場合は、一次ボリュームへのその後の更新を防止することができる。このため、リアルタイムの完全エラー災害復旧が達成される。

【0034】非同期データ・シャドーイング非同期遠隔データ・シャドーイングは、1回の災害で一次側と二次側の両方が増悪してしまう確率を低減させたために一次側と二次側の距離をさらに大きくする必要がある場合、または一次側アプリケーションのパフォーマンスへの影響を最小限に拘束する必要がある場合に使用する。一次側と二次側との距離は、現在では地球全体またはそれ以上に延長できるが、複数の一次サブシステムは背後にある複数のDASDボリュームにわたる書き込み更新を複数の二次サブシステムに同期させることは、さらに複雑である。二次記憶サブシステム上でシャドーイングするために、一次データ・ムーバを介して一次記憶制御装置から二次データ・ムーバ・レベコデータ更新を御送することができ、両者間でやりとりされる制御データの量は、最小限でなければならず、同時に、複数の記憶制御装置に隠れている複数のDASDボリュームにわたる二次システム上でレコデータ書き込み更新の順序を正確に再構築できるものでなければならぬ。

【0035】図4は、一次側421と遠隔側または二次側431とを含む非同期災害復旧システム400を示している。一次側421は、DFSMS/MVSホスト・ソフトウェアを実行するIBM ES/9000などの一次プロセッサ401を含む。一次プロセッサ401

この通信リンク408は、電話(T1、T3回線)、無線、無線/電話、マイクロ波、衛星などの複数の適当な通信方式によって実現できる。

【0039】非同期データ・シャドーイング・システム400は、一次DASD406へのすべてのデータ書き込みの順序が保持され、二次DASD416に適用される(すべての一次記憶サブシステムにわたるデータ書き込み順序を保持する)ように、一次記憶制御装置405から制御データを収集する機能を含む。二次側431に送られるデータおよび制御情報は、データ保全性を保持するのに一次側421の存在が重要になるほど、十分なものでなければならぬ。

【0040】アプリケーション402、403は、データまたはレコード更新を生成するが、このレコード更新は、一次記憶制御装置405によって収集され、PDM404によって読み取られる。それぞれの一次記憶制御装置405は、非同期遠隔データ・シャドーイング・セッションのためにそれぞれのレコード更新をグループ化し、非特定一次DASD406のREAD要求を紹介する。PDM404にこれらのレコード更新を提供する。一次記憶制御装置405からPDM404へのレコード更新の転送は、START入出力操作の回数を最小限にしながら、各一次記憶制御装置405と一次プロセッサ401との間で転送されるデータの量を最大にするように、PDM404によって制御され、最適化される。PDM404は、非特定READ間の時間間隔を変えることで、一次記憶制御装置405とこの最適化だけでなく、二次DASD416用のレコード更新の適用期間も制御することができる。

【0041】データ保全性を維持しながら、PDM404がレコード更新を収集し、そのレコード更新をSDM414に送信するには、すべての一次記憶サブシステムにおいて二次DASD416に対して行われる一次DASD406のレコードWRITEシーケンスを再構築するに十分な制御データとともに、特定の期間の間、適切な複数の時間間隔でレコード更新を送信する必要がある。一次DASD406のレコードWRITEシーケンスの再構築は、自己記述レコードをPDM404からSDM414に渡すことによって達成される。SDM414は、所与の時間間隔分のレコードが紛失しているかどうか、または不完全になっているかどうかを判断するために、その自己記述レコードを検査する。

【0042】図5および図6は、接頭部ヘッダ500(図5)と、一次記憶制御装置405によって生成されたレコード・セット情報600(図6)とを含む、各自記述レコードごとにPDM404が作成するジャーナル・レコード形式を示している。各自記述レコード416は、それぞれの時間間隔の時間順に二次DASD416に適用できるように、それぞれの時間間隔ごとにさらにSDM414によってジャーナル処理される。

は、IMSおよびDBSアプリケーションなどのアプリケーション・プログラム402および403と、一次プロセッサ(PDM)404をさらに含む。一次プロセッサ401には、そこで実行されるすべてのアプリケーション(402、403)に共通の基準を提供するために、共通シスプレックス・クロック(synplex clock)407が設けられ、すべてのシステム・クロックまたは時間隔(図示せず)がシスプレックス・クロック407に同期し、すべての時間依存プロセスが相互に正しいタイミングで動作するようになっている。たとえば、一次記憶制御装置405は、単一の一次記憶制御装置406への2回の連続する書き込み出力操作が同じタイム・スタンプ量を示さないように、複数のレコード書き込み更新時間を確実に区別するのに適した解像度に同期している。シスプレックス・クロック407の解像度(正確さではない)は重要である。PDM404は、シスプレックス・クロック407に接続された状態で図示されているが、書き込み出力操作がそこで発生するわけではない。また、一次プロセッサ401が単一の時間基準(たとえば、単一のマルチプロセッサES/9000システム)を有する場合に、シスプレックス・クロック407は不要である。

【0036】一次プロセッサ401には、IBM3990-6型記憶制御装置などの複数の一次記憶制御装置405が光ファイバ・チャネルなどの複数のチャネルを介して接続されている。また、各一次記憶制御装置405には、IBM3990DASDなどの複数の一次DASD406からなる少なくとも1つのストリングが接続されている。一次記憶制御装置405と一次DASD406によって、一次記憶サブシステムが形成される。各記憶制御装置405と一次DASD406は、個別のユニットである必要はなく、両者を組み合わせて単一のドロブにしてもよい。

【0037】一次側421から数千キロメートル離れた位置に配置される二次側431は、一次側421と同様に、そこで動作する二次データ・ムベ(PDM)416を有する二次プロセッサ411を含む。二次プロセッサ411には、当技術分野で既知の通り、光ファイバ・チャネルなどのチャネルを介して複数の二次記憶制御装置415が接続されている。記憶制御装置415には、複数の二次DASD416と1つの制御情報DASD417(複数の可)が接続されている。記憶制御装置415とDASD416および417によって、二次記憶サブシステムが構成される。

【0038】一次側421は、通信リンク408を介して二次側431と通信する。より具体的には、一次プロセッサ401は、仮記憶通信アクセス方式(VTAM)通信リンク408などの通信プロトコルによって、二次プロセッサ411にデータと制御情報を転送する。

【0043】ここで図5を参照すると、各レコード・セットの先頭に挿入される接頭部ヘッダ500は、接頭部ヘッダ500と、各レコード・セットごとにSDM414に送信される実際の一次レコード・セット情報600との長さの合計を記述するための総データ長さ501を含む。操作タイム・スタンプ502は、PDM404が現在処理している操作セットの開始時間を示すタイム・スタンプである。この操作タイム・スタンプ502は、1組の一次記憶制御装置405に対してREAD RECORD SET機能を実行する際に(シスプレックス・クロック407に応じて)PDM404によって生成される。一次DASD406の書き込みの入力時間610(図6)は、各一次記憶制御装置405のREAD RECORD SETごとに固有のものである。操作タイム・スタンプ502は、すべての記憶制御装置で共通のものである。READ RECORD SETコマンドは、PDM404によって出されるが、以下の条件のいずれかの場合に予測できる。

(1) 一次記憶制御装置405の所定のしきい値に基づく、その一次記憶制御装置のアランジョン制込み  
(2) 所定の時間間隔に基づく、一次プロセッサ401のタイマ制込み

(3) レコード・セット情報が、使用可能であるがまだ読み取られていない未解決のレコード・セットに関する追加情報を示す場合

条件(2)では、タイマ間隔を使用して、低レベル活動の期間中に、アクセス431がどの程度遅れて実行するかを制御する。条件(3)は、PDM404が一次記憶制御装置405の活動に遅れないようにするためにさらに活動を駆動する処理期間中に、PDM404がすべてのレコード・セットを待ち行列処理しなかった場合に発生する。

【0044】時間間隔グループ番号503は、現行レコード・セット(整合性グループのうちの所与の時間間隔グループ)についてすべての一次記憶制御装置405にわたるレコード(セット)が属する時間間隔(操作タイム・スタンプ502とレコード総取り時間507によって境界が示される)を識別するためにPDM404が出力する。グループ内順序番号504は、所与の時間間隔グループ503内の各レコード・セットごとに一次記憶制御装置405用のアプリケーションWRITE入出力の順序を(PDM404に対して)識別するためにヘッダウェアが提供するIDである。一次SSID(補助記憶ID)505は、各レコード・セットごとに一次記憶制御装置405の特定の一次記憶制御装置を明確に識別するものである。二次ターゲット・ポリシー506は、パフォーマンス上の考慮事項に応じて、PDM404またはSDM414のいずれかに基づいて割り当てられる。レコード総取り時間507は、すべての一次記憶制御装置405に共通の操作タイム・スタンプを提供し、現行

間隔のレコード・セットの終了時間を示す。

【0045】次に図8を参照して説明すると、レコード・セット情報600は、一次記憶制御装置405によって生成され、PDM404によって収集される。更新間隔情報601〜610は、レコード更新が行われた実際の一次DASD406を含む各レコードの一次装置ユニット・アドレス601を含む。レコード番号/ヘッダ番号(CCHH)602は、各レコード更新ごとの一次DASD406上の位置を示す。一次記憶制御装置のセットシヨニDである二次SSID603は、一次SSID505と同じものである。状況フラグ604は、特定のデータ・レコード620が後に続くかどうかに関する状況情報を提供する。順序番号605および630は、レコード・セット全体(PDM404に転送されたすべてのデータ)が読み取られたかどうかを示すために各レコードに番号を1つずつ割り当てる。一次DASD書き込み出力タイプ606は、各レコードについて行われた書き込み操作のタイプを識別する操作標準である。この操作標準は、更新済み、フォーマット済み、部分ドロップ・レコード・フォロー、完全トラック・データ・フォロー、消去コマンド実行、または全書き込み実行を含む。検索索引数607は、最初に読み取られたレコード・セット・データ・レコード620に関する初期位置決め情報を示す。セクタ番号608は、レコードが更新されたセクタを識別する。カウント・フィールド609は、後続の特定のレコード・データ・フィールド620の数を記述する。一次DASD406の書き込み更新が行われたホスト・アプリケーション時間は、更新時間610に記録される。特定のレコード・データ620は、各レコード更新ごとのカウント/キー/データ(CKD)フィールドを提供する。最後に、順序番号630は、読み取られたレコード・セット全体がPDM404に転送されたかどうかを示すために順序番号605と比較される。

【0046】一次DASD406でレコード更新が書き込まれたのと同じ順序でSDM414がそのレコード更新をコピーできるように、ソフトウェア・グループが呼び出した整合性グループで更新レコードが処理される。整合性グループを作成するために使用する情報(すべての記憶制御装置405から収集したすべてのレコード・セットにわたる)は、操作タイム・スタンプ502、時間間隔グループ番号503、グループ内順序番号504、一次制御装置SSID505、レコード総取り時間507、一次装置アドレス601、一次SSID603、および状況フラグ604を含む。1つの時間間隔グループのすべてのレコードがSDM414側で各記憶制御装置405ごとに受け取られたかどうかを判断するために使用する情報は、時間間隔グループ番号503、グループ内順序番号504、物理制御装置ID505、および一次SSID603を含む。完全復旧可能な一次DASD406のレコード更新と同等に二次DASD416上に

レコード更新を配置するのに必要な情報は、二次データ・グループ・ボリューム506、CCHH602、一次DASデータ読み出し/タイパ606、検索引数607、セクタ番号608、カウンタ609、更新時間610、および特定のレコード・データ620を含む。

【0047】図7および図8は、復旧時間とジャーナル転送時間を単純化した現行ジャーナル内容を記述するための状態テーブル700とマスター・ジャーナル800をそれぞれ示す。状態テーブル700は、PDM404またはSDM414のいずれかが収集した構成情報を提供し、一次記憶制御装置のセッションID (SSID番号) およびその制御装置でのボリュームと、対応する二次記憶制御装置のセッションIDおよび対応するボリュームとを含む。このため、構成情報は、どの一次ボリューム710または一次DASDエクス Tent が二次ボリューム711または二次DASDエクス Tent にマッピングされるかを追跡する。状態テーブル700まで単純に拡張して部分ボリューム・エクス Tent 712 (CCHHからCCHHまで) を示す場合、部分ボリューム識別コードは、ここに記載するのと同じ非同期遠隔コピー方法を使用して達成できるが、完全ボリュームの場合より細分性 (トラックまたはエクス Tent) はより細かくなる。

【0048】マスター・ジャーナル800は、整合性グループ番号、ジャーナル・ボリューム上の位置、および操作タイム・スタンプを含む。また、マスター・ジャーナル800は、整合性グループにグループ化された特定のレコード更新を維持する。状態テーブル700とマスター・ジャーナル800は、災害復旧をサポートするため、一次システム410がもはや存在しないスタンバイアロン環境で動作できなければならない。

【0049】制御項目全体が正しく書き込まれるようにするため、タイム・スタンプ制御は各マスター・ジャーナル800の前後に置かれる。このタイム・スタンプ制御は、さらに二次DASD417に書き込まれる。制御要素は二重項目 (1) および (2) を含み、次に示す例のように一方の項目が必ず実行項目になる。

(1) タイム・スタンプ制御 | 制御情報 | タイム・スタンプ制御

(2) タイム・スタンプ制御 | 制御情報 | タイム・スタンプ制御

いかなる時点でも、(1) または (2) のいずれかの項目が実行または有効項目になるが、有効項目は前後に等しいタイム・スタンプ制御を持つ項目である。災害復旧では、制御情報を得るために、最新のタイム・スタンプを持つ有効項目を使用する。この制御情報は、状態情報 (記憶制御装置、装置、および適用される整合性グループに関する緊密情報) とともに、二次記憶制御装置415にどのレコード更新が適用されたかを判別するのに使用する。

【0050】整合性グループ所定の時間間隔の間にすべての一次記憶制御装置405にわたるすべての読取りレコード・セットが二次側431で受け取られ、SDM414は、受け取った制御情報を解釈し、レコード更新が最初に受け取ったDASD406上で書き込まれたのと同じ順序でそのレコード更新が適用されるように、受け取った読取りレコード・セットをレコード更新・グループとして二次DASD416に適用する。このため、一次側がジャーナルの順序 (データ保全性) 整合性はすべて二次側431で維持される。このプロセスは、以下、整合性グループの形成と呼ぶ。整合性グループの形成は、次のような仮定に基づいて行われる。(A) 独立しているアプリケーション書き込みが制御装置の順序命令に違反しない場合は、どのような順序でもアプリケーション書き込みを実行できる。

(B) 従属しているアプリケーション書き込みは、タイム・スタンプの順に実行しなければならないため、アプリケーションは、書き込み番号1から制御装置終了、装置終了を受け取る前に従属書き込み番号2を実行することができない。(C) 第二の書き込みは必ず (1) 遅いタイム・スタンプを持つ第一の書き込みと同じレコード・セット整合性グループに入るか、(2) 後続のレコード・セット整合性グループに入る。

【0051】図9を参照して説明すると、図中には、記憶制御装置SSID1、SSID2、およびSSID3など (記憶制御装置はいくつかも含まれることができる) が、この例では明解にするため3つ使用する) に関する整合性グループの形成例 (整合性グループは一次側421または二次側431のいずれかに形成できる) があり、1または2は、時間間隔T1、T2、およびT3が示される。時間間隔T1、T2、およびT3は昇順に発生するものと想定する。時間間隔T1の操作タイム・スタンプ502は、記憶制御装置SSID1、SSID2、およびSSID3について設定されている。PDM404は、時間間隔T1〜T3の間に記憶制御装置SSID1、2、および3からレコード・セット・データを入手する。時間間隔T1のSSID1、2、および3に関するレコード・セットは、時間間隔グループ1であるG1 (時間間隔グループ番号503) に割り当てられる。グループ内順序番号504は、SSID1、2、および3のそれぞれについて示され、この場合、SSID1は11:59、12:00、および12:01に3つの更新を持ち、SSID2は12:00および12:02に2つの更新を持ち、SSID3は11:58、11:59、および12:02に3つの更新を持つ。時間間隔T2およびT3のレコード・セットは列挙されているが、簡略化のため、更新時間の例は示されていない。

【0052】ここで、二次側431で受け取った制御情報およびレコード更新に基づいて、整合性グループNを生成することができる。時間間隔グループ番号1のレコ

ード更新が時間間隔グループ番号2のレコード更新より遅くならないようにするため、記憶制御装置SSID1、2、および3のそれぞれの最後のレコード更新の最も早い読取りレコード・セット時間と等しい、最小時間が設定される。この例では、最小時間は12:01になる。最小時間と等しいかそれ以上の読取りレコード・セット時間を有するレコード更新はすべて整合性グループN+1に含まれる。1つのポリシーとして2つのレコード更新時間が等しい場合、システムックス・クロック407の1/16分等級度与えられる可能性はほとんどないが、時間間隔グループN内の早い順序番号を持つレコード更新は、整合性グループN用のそのグループとともに保管される。ここで、レコード更新は、読取りレコード・セット時間に基づいて順序づけされる。複数のレコード更新の時間が等しい場合、小さい順序番号を持つレコード更新は、大きい順序番号を持つレコード更新の前に置かれる。これに対して、複数のレコード更新のタイム・スタンプが等しいが、ポリシーが異なる場合は、そのレコード更新が同じ整合性グループに保管されている限り、任意の順序にすることができ。

【0053】一次記憶制御装置405が、指定の時間間隔の間に読取りレコード・セットへの応答を完了しなかった場合、その一次記憶制御装置405が完了するまで、整合性グループを形成することはできない。一次記憶制御装置405がその操作を完了しなかった場合は、未着制込みのために、システムの未着制込みハンドラが制御権を受け取り、操作が終了する。これに対して、一次記憶制御装置405が適切な時間に操作を完了した場合は、入出力が完了に至り、通常操作が実行される。整合性グループの形成では、一次記憶制御装置405に対する書き込み操作がタイム・スタンプが付けられると予想される。しかし、プログラムによっては、タイム・スタンプが付けられずに書き込みが生成されるものもある。この場合、一次記憶制御装置405は、タイム・スタンプとしてゼロを返す。整合性グループの形成は、データが読み取られたタイム・スタンプに基づいて、タイム・スタンプを持たないこれらのレコードの境界を示すことができる。整合性グループの時間別にレコード更新の境界を容易に示せないほど、タイム・スタンプを持たないレコード更新が一定の時間間隔の間に多数発生した場合、二重ボリュームが同期していないというエラーが発生する可能性がある。

【0054】図10および図11は、整合性グループを形成する方法を示す流れ図である。図10を参照して説明すると、このプロセスは、ステップ1000から始まり、一次側421が、行うべき遠隔データ・シャドローンを確立する。ステップ1010では、システムックス・クロック407を同期クロック (図4) として使用して、すべてのアプリケーション入出力操作にタイム・スタンプが付けられる。PDM404は、ステップ1050

【0055】ステップ1040は、前述の通り、アテンション・メッセージを含むプロンプト、所定のタイム・間隔、または読取りレコード・セット情報600を各一次記憶制御装置405から読み取ることを含む。ステップ1050でPDM404がレコード・セットの読取りを開始すると、PDM404は、各レコード・セットの前に特定のジャーナル・レコード (ジャーナル・レコードは、接頭部ヘッダ500と、レコード・セット情報600を含む) を作成するための接頭部ヘッダ500 (図5を参照) を付ける。このジャーナル・レコードには、二次側431 (または一次側421) で整合性グループを形成するのに必要な制御情報 (およびレコード) が含まれる。

【0056】ステップ1060では、PDM404が通信リンク408を介して整合性グループがそこで作成される場合は、同じデータ・ムーブ・システム内で) SDM414に生成したジャーナル・レコードを送信する。SDM414は、ステップ1070で状態テーブル700を使用し、データ・シャドローイング・セッションに確立した各時間間隔グループおよび一次記憶制御装置405ごとに、受け取ったレコード更新をグループ番号および順序番号別に収集する。ステップ1080でSDM414は、ジャーナル・レコードを検査し、各時間間隔グループごとにすべてのレコード情報を受け取ったかどうかを判別する。ジャーナル・レコードが不完全な場合は、ステップ1085によって、SDM414はPDM404に必要なレコード・セットを再送信するように通知する。PDM404が正しく再送信できない場合は、二重ボリューム時に障害が発生している。ジャーナル・レコードが完全な場合は、SDM414による整合性グループの形成を含むステップ1090が実行される。

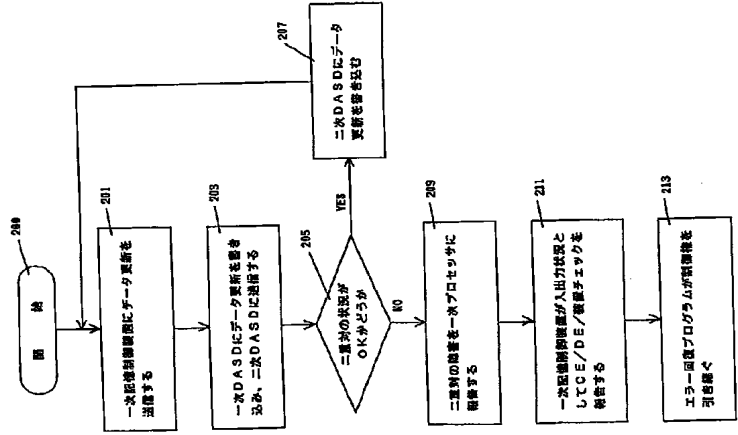
【0057】図11を参照すると、図中には、整合性グループを形成するためのステップ1090 (図10) を表すステップ1100〜1160が示されている。整合性グループの形成は、ステップ1100から始まるが、このステップでは、各ソフトウェア整合性グループが二次DASD417 (図4) 上のSDM414ジャーナル・ログ (hardened) に書き込まれる。ステップ1110は、時間間隔グループが完全かどうかを判別するテストを実行する。すなわち、各一次記憶制御装置405は、少なくとも1つの読取りレコード・セット・パツ



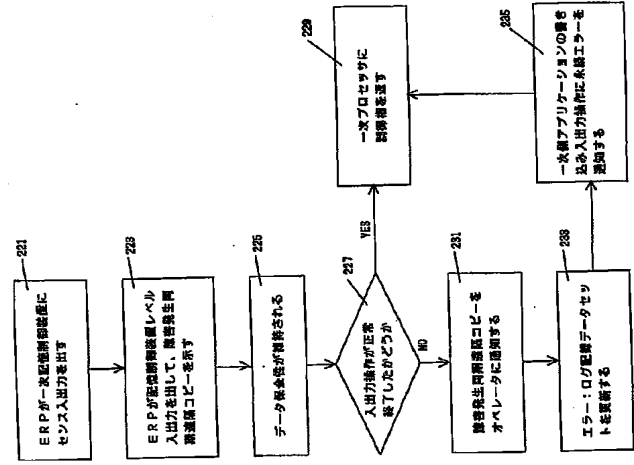




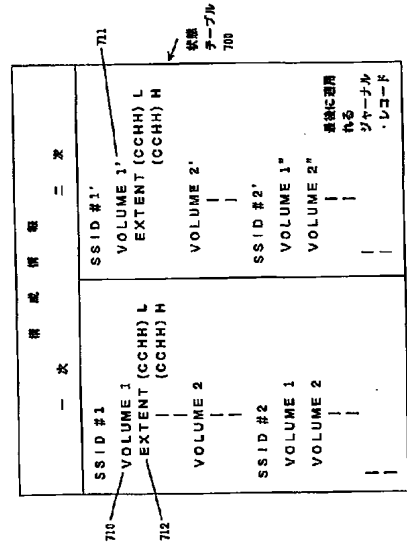
【図2】



【図3】



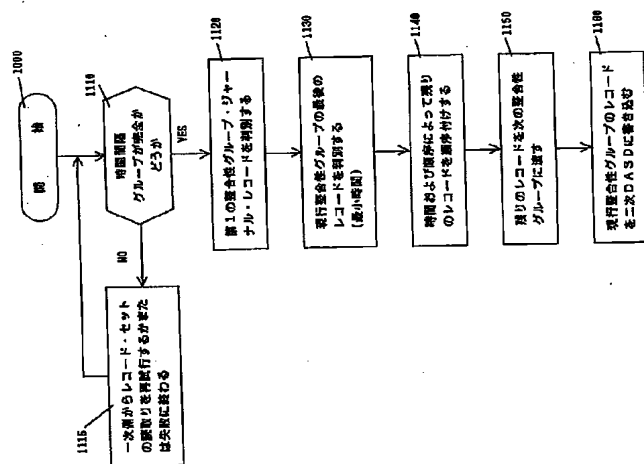
【図7】







【图11】

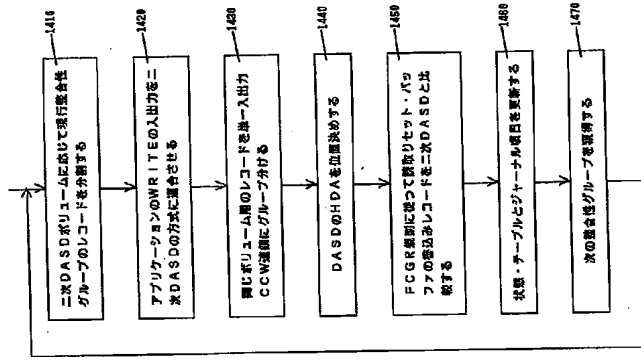


【图12】

[illegible]



【図16】



フロントページの続き

- |         |                       |         |                       |
|---------|-----------------------|---------|-----------------------|
| (72)発明者 | ウィリアム・フランク・ミツカ        | (72)発明者 | マイケル・アロイシウス・ポールソン     |
|         | アメリカ合衆国85718 アリゾナ州ツー  |         | アメリカ合衆国95037 カリフォルニア州 |
|         | ン イースト・ラズベラダ 3921     |         | モーガン・ヒル ホウィツバーク・ド     |
| (72)発明者 | クラウス・ウィリアム・ミツケルセン     |         | ライプ2901               |
|         | アメリカ合衆国95120 カリフォルニア州 | (72)発明者 | ロバート・ウェズリー・ショムラー      |
|         | サンノゼ シーナリー・コート・ドライ    |         | アメリカ合衆国95037 カリフォルニア州 |
|         | 6557                  |         | モーガン・ヒル ピエドメント・コート    |
|         |                       |         | 17015                 |